Wearable Common Sense:
Gisting Conversations with Contextual Information and Common Sense

Nathan Eagle, Push Singh, Alex (Sandy) Pentland
MIT Media Lab
nathan@media.mit.edu, push@media.mit.edu, sandy@media.mit.edu

Abstract

This paper introduces a system that incorporates both contextual and commonsensical information to understand the gist of an informal, face-to-face conversation. We show that wearable devices, such as PDAs or cell phones, can provide the valuable contextual information critical for robust classification of a detailed conversation topic.

1. Motivation

Once wearable computers are able to capture the gist of a user's conversation, an enormous number of potential applications become possible. However, current topic-spotting methods have met with little success in characterizing spontaneous conversations involving hundreds of potential topics [3]. This paper claims that performance can be greatly improved by making use of not only the text of a transcription, but also contextual and commonsensical information from the dialogue.

Why is this problem hard? Even with the latest speech recognition engines trained to a user's voice, accuracy rates for spontaneous speech recognition fall below 35%. Conversation transcripts like the one shown below are difficult even for a human to comprehend.

> Store going to stop and listen to type of its cellular and fries he backed a bill in the one everyone get a guess but that some of the past like a salad bar and some offense militias cambers the site fast food them and the styrofoam large chicken nuggets son is a pretty pleased even guess I as long as can't you don't have to wait too long its complicity sunrise against NAFTA pact if for lunch

Additional context, such as information that the conversation is occurring in a cafeteria, or more precisely, waiting in line to be served, would help many people understand that the user is talking about what to get for lunch. Prior knowledge about the conversation participants and the time of day may also significantly augment a person's ability to infer the gist of the interaction. Our work suggests that the additional contextual and commonsensical information that a human can employ for inference on the transcript above is equally helpful to a probabilistic model.

We have enabled a suite of wearable computers with the ability to provide the contextual information necessary for a human to infer a conversation's gist. To take the human fully out of the loop, the next step is to infuse the system with commonsense knowledge, for example: people typically eat lunch at noon in cafeterias, or that eating is often motivated by feeling hungry. Using a database of commonsense knowledge [2] combined with contextual information gleaned from the wearable computers, we show that gisting conversations can indeed be tractable.

2. Implementation

A system of linux-based PDAs, the Sharp Zaurus, and a variety of both wireless (Bluetooth) and standard wired cellular phone headsets were used for preliminary data collection. The PDAs were equipped with CF 802.11b wireless cards that capture access point information and stream audio to a central sever. We use the OMCSNet semantic network, containing over 250,000 commonsensical semantic relationships contributed from over 10,000 people across the web [2, 4]. OMCSNet uses a hybrid knowledge representation strategy where individual concepts are represented linguistically, and are related by a small set of about twenty specific semantic relationships such as LocationOf, SubeventOf, and HasEffect. Despite this vast amount of data, the knowledge database is less than 50 MB, well within the 256 MB limit of the SD card in the PDAs.

A TCL script was written to interface with the SDK of the commercial speech recognition engine ViaVoice. The script inputs words directly into our system for pre-processing consisting of lemmatizing, removing stop words (such as 'the', 'like', etc), and then semantic filtering. While the words the speech recognition engine gets correct tend to be grouped around neighboring semantically-related nodes in OMCSNet, errors in the transcriptions turn out to be distributed randomly over this network. A clustering technique is employed on the words to generate a list of keywords and the nodes they described are assumed to be potentially indicative of a user's conversation. These selected nodes are then input into the probabilistic model and are weighted by their number of supporting keywords [1].

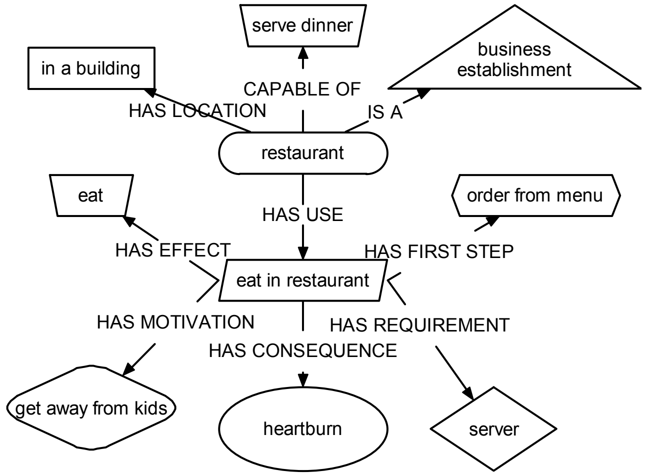

Figure 1. A Selection of OMCSNet's 250,000 Relations

By flattening the networks of the different relationship types, a bipartite network can be formed to incorporate all ties from words to gists. The probability of a specific gist can be modeled as proportional to the sum of a gist's links to the selected words and any available supporting contextual information:

Gist without Context:

$$P(g_i|k) = \frac{k_i}{\sum_{i=1}^{G} k_i}$$

Gist with Context from Wearable:

$$P(g_i|k,c) = \frac{k_i + c_i}{\sum_{i=1}^{G} (k_i + c_i)}$$

where k_i is the number of links between a gist, g_i, and the transcript keywords. G is the number of potential gists (approximately 700) and c_i is the number of links between the context and the keywords.

3. Results

The additional contextual information from the wearable dramatically increases the confidence scores of the classifier. Because the conversation was being streamed to an access point ID mapped to 'cafeteria', this information was also passed to the server along with the audio. Using the location as a bias, the confidence of the classifier becomes much more robust:

Without Location Context		**With Location Context**	
5	buy hamburger	27	eat in fast food restaurant
5	eat in fast food restaurant	21	eat in restaurant
5	eat in restaurant	18	wait on table
5	talk with someone far away	16	you would go to restaurant because you
5	buy beer	16	wait table
4	go to hairdresser	16	go to restaurant
4	wait in line	15	know how much you owe restaurant
4	howl with laughter	12	store food for people to purchase
4	eat healthily	11	sitting down while place order at bar
4	play harp	11	cook food

Table 1. Confidence Scores using Context for Gist Differentiation

4. Conclusions

The robustness of the classifier comes from its ability to bias the prior probability of each node based on other contextual information from the wearable, such as the user's location, conversation participants, or simply the people in his local proximity. Online learning algorithms are being developed to incorporate subsequent observations into the classifier to augment the commonsense model with more a specialized model that better reflects an individual's behavior.

As wearable computers become ever more embedded in society, the additional contextual information they provide about a user's context will become invaluable for a variety of applications. This paper has shown the potential for these devices to leverage this additional information to begin understanding informal face-to-face conversations.

5. Acknowledgments

This work was partially supported by the NSF Center for Bits and Atoms (NSF CCR-0122419).

6. References

[1] Eagle, N., Singh, P., Pentland, S. Common Sense Conversations, to be published in the *Artificial Intelligence, Information Access, and Mobile Computing Workshop at the 18th International Joint Conference on Artificial Intelligence* (IJCAI). Acapulco, Mexico, 2003.

[2] Liu, H. and Singh. P. OMCSNet: A commonsense inference toolkit. In submission. 2003.

[3] Jebara, T., Ivanov, Y., Rahimi, A., Pentland, A. Tracking Conversational Context for Machine Mediation of

Human Discourse. *Proceedings of AAAI Fall Symposium on Socially Intelligent Agents* : AAAI. 2000.

[4] Singh, P. The public acquisition of commonsense knowledge. *Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access.* Palo Alto, CA: AAAI. 2002.